Košuta Estera Lerga

Faculty of Humanities and Social Sciences, University of Rijeka, Croatia
kosuta.lerga@gmail.com

Lucia Načinović Prskalo, Marija Brkić Bakarić

Faculty of Informatics and Digital Technologies, University of Rijeka, Croatia
lnacinovic@uniri.hr, mbrkic@uniri.hr

Adapting the Generic English-Croatian NMT Model to a Religious Domain

Abstract

Recent discoveries in the field of artificial intelligence have significantly impacted various professions, including the translation industry, leading to notable changes in translators' work processes. The study presented in this article indicates that today any translator, even those without advanced IT skills, can develop a higher quality Neural Machine Translation (NMT) system based on their own texts. This paper evaluates Google's AutoML Translation service, which enables users to train high-quality models using their own text data. Specifically, AutoML Translation integrates an additional layer that tailors the generic Translation API model to a specific domain. The training process involves providing a user-defined dataset containing aligned sentences in the source and target languages. Google's AutoML Translation service was used to adapt the base English-Croatian Google NMT model to the field of religion. Following a brief introduction to machine translation, this paper outlines the key aspects of the training and evaluation processes. Additionally, it presents two corpora employed in the training phase. The results demonstrate that a customized model outperforms the base model, as evidenced by the BLEU score.

Keywords: automatic translation, domain adaptation, neural machine translation, religious domain, aligned parallel corpora

1.Introduction

When presented with a message, there are numerous methods to effectively communicate its intended meaning. Similarly, it is highly probable that any translator among a group would offer slightly varied translations of a message originally conveyed in the source language. The inherent subjectivity of this task contributes to the inherent complexity associated with machine translation (MT) and its subsequent evaluation.

MT approaches generally fall into two categories: rule-based methods and data-driven approaches. Rule-based methods were predominant before the 2000s,

characterized by their subjective and labour-intensive nature, making them susceptible to unforeseen language phenomena and scalability issues. Rule-based systems involve linguists crafting specific rules to transform source language into target language. In contrast, data-driven approaches emerged in the 1980s, aiming to learn translation patterns by analysing numerous pairs of human-translated segments. Data-driven MT encompasses example-based MT, statistical machine translation (SMT), and neural machine translation (NMT). Example-based MT retrieves similar examples from pairs of human-translated sentences to generate translations. The concept of SMT originated in the late 1980s but gained mainstream acceptance around 2000. Neural network approaches gradually integrated into various components of SMT, reaching their full potential from 2015 onwards. Despite their existence in the previous century, the computational complexity associated with these methods hindered serious advancements beyond toy examples (Koehn 2020). Sequence-to-sequence models eventually replaced traditional phrase-based approaches in NMT systems based on the encoder-decoder paradigm (Chen et al. 2018). The first production Google's NMT system was presented by Wu et al. (2016).

Comparative quality analyses of neural machine translation systems versus statistical machine translation systems, as detailed by Koehn and Knowles (2017), indicate that neural machine translation systems often achieve lower quality on out-of-domain texts, favouring fluency over adequacy to a point of sacrificing the latter. Consequently, they are more sensitive to domain mismatches compared to SMT systems (Ruopp 2020). Furthermore, these models exhibit a "steeper" learning curve concerning data volume, resulting in reduced performance in low-resource settings (Koehn and Knowles 2017).

The datasets utilized in this research can serve valuable pedagogical purposes in training translation students and as foundational resources for interdisciplinary research in fields such as translation studies, cross-linguistic analysis, and lexical semantics. Moreover, they facilitate the refinement of large language models and the adaptation of existing machine translation services. In this study, we leverage these datasets to adapt the base English-Croatian Google NMT model to the domain of religion using Google's AutoML Translation service.

The organization of the paper is as follows: Section 2 briefly presents related work. Section 3 describes the training procedure and provides details about the two aligned parallel corpora used in this process. The results and discussion follow, which are then summarized in the concluding remarks presented in the final section of the paper.

2. Related Work

Carlson et al. (2018) utilized various versions of the Bible aligned by chapter and verse numbers to create a corpus for the style transfer task. Style transfer can be viewed as a form of monolingual translation, akin to a machine translation prob-

lem where the source and target languages differ only in terms of style. The authors incorporated thirty-four stylistically distinct Bible versions, including the archaic language of the King James Version, which dates back centuries. The study involved training and evaluating both an encoder-decoder recurrent neural network and an SMT system. The neural system outperformed the statistical approach when more substantial changes are required to the source segment.

Viswanathan et al. (2019) used Google's AutoML Translate to train a system aimed at producing more consistent translations concerning register, specifically tone and style, while still harnessing the capabilities of a general-purpose MT system. The authors specifically focused on the register associated with personal pronouns. The task can be viewed as a special case of domain adaptation. To accomplish this, they employed formality-specific datasets to train custom models that are strongly biased towards the respective registers. They repeated the training procedure multiple times on the same training dataset, replacing the model with the one obtained from the previous iteration, using Google's generic NMT model as the base model. The results indicate that fewer than 5000 sentences may be sufficient to leverage transfer learning effectively from the base model.

Ruopp (2020) utilized Google AutoML Translation to train custom NMT engine adapted to COVID-19. The study acknowledges the importance of the translation memory format in the era of adaptive, document-context aware NMT systems for preserving document context. Significant improvements in BLEU scores were achieved. However, the author also highlights the risks associated with domain adaptation for high resource language pairs, as adapting to one domain can lead to a deterioration in quality even for closely related domains. Furthermore, combining training data from different sources may blend translator-specific or organization-specific preferences embedded in the training data due to quality assurance procedures, potentially resulting in contradictions.

AutoML Translation is also employed to customize Google Translate for three different genres: song lyrics, novels, and subtitles. Higher BLEU scores are reported in all three cases, with the most significant increase in BLEU observed for subtitles (Al-Sabbagh 2024).

3. Research Study

The primary focus of this paper is to explore how Google's AutoML Translation service enables translators, including those with limited IT expertise, to create improved Neural Machine Translation (NMT) systems using their own text data. Additionally, this study aims to analyse the impact on the workflow within the translation industry, considering factors such as the BLEU score and subjective perceptions of translation quality. As this study focuses on translating religious texts, the initial step involved acquiring aligned parallel corpora of texts within the religious domain. Descriptions of the dataset are provided in the following subsection.

3.1. Data Description

The primary focus of this research centres around the translation of religious texts, necessitating the initial step of acquiring a well-matched parallel corpus of texts within this domain. The corpora used in the training procedure include selected texts by William Branham and their respective translations, as well as the King James Version of the Bible and its translation by Ivan Vrtarić from 2016.

The first corpus consists of texts in English authored by William Branham and their translations. Each text was translated by a single translator, although contributions were made by several different translators. Two versions of the dataset are utilized—one aligned based on paragraphs (BranhamTextsPars) and another based on sentences (BranhamTextsSents). Two randomly extracted excerpts are provided in Tables 1 and 2, respectively, offering insights into the translation alignment process and the structure of the dataset.

Table 1. Selected excerpts from the paragraph-level aligned corpus of William Branham texts (BT_{Pars})

Gracious Lord, we bring to Thee these little parcels, perhaps some of them look to be maybe little vests for a baby, or--or some little undershirt, or maybe a little pair of booties, or--or something, a handkerchief, that's going to the sick and the afflicted. Lord, it is according to Thy Word that we do this. For we read, in the Book of Acts, that they taken from the body of Your servant, Paul, handkerchiefs and aprons, because they believed that Your Spirit was on the man. And unclean spirits went out of people, and afflictions and diseases left them, because they believed. And now we realize, Lord, that we're not Saint Paul, but we know that You still remain Jesus. And we pray that You'll honor the faith of these people.	Dragi Gospode, donosimo ti ove male komadiće materijala, možda neki sliči na mala odijelca za bebu, ili - ili neku malu potkošulju ili možda mali par cipelica, ili - ili tako nešto, maramice koje će ići k bolesnima i napaćenima. Gospode, to činimo prema tvojoj Riječi, jer čitamo u Djelima Apostolskim da su uzimali maramice i ubruse s tijela tvoga sluge Pavla, zato što su vjerovali da je tvoj Duh bio na tom čovjeku. I nečisti duhovi su izlazili iz ljudi, a bolesti i muke su ih napuštale, zato što su vjerovali. A sada mi shvaćamo Gospode da mi nismo sveti Pavao, ali znamo da si ti i dalje Isus. I molimo da ti poštuješ vjeru ovih ljudi.
Not at the first day: dead form; second day there was a rumor (See?)--or the second day it was still dead: Luther, Wesley. At the beginning of the third day there was a rumor around. Nothing on the first day (Luther), nothing on the second day, and in the third day (the dispensation of the Holy Spirit) there was a rumor around that He was the same yesterday, today, and forever. But at the end of the third day, that's where He made Hisself known, come right among them, come among His people and said, «Look at Me; I'm the same One.»	Ne prvog dana, tada je bila mrtva forma; drugog dana bila je glasina - drugog dana još je uvijek bilo mrtvo - Luther, Wesley. Na početku trećeg dana pojavila se glasina. Ništa se nije dogodilo prvog dana (Luther), ništa se nije dogodilo ni drugog dana, a trećeg dana (etapa Duha Svetog) se pojavila glasina da je On isti jučer, danas i zauvijek će biti isti. Ali na kraju trećeg dana Se On obznanio. Došao je među njih, među Svoje ljude i rekao: «Pogledajte Me, Isti Sam.»

Table 2. Selected excerpts from the sentence-level aligned corpus of William Branham texts (BT_{Sents})

All down through the ages they received the Holy Spirit, but not in the measure that they have It now; 'cause it's a restoration of the first.	Skroz kroz doba oni su primali Duha Svetoga, ali ne u mjeri u kojoj Ga sada imaju jer je obnova prvog.
Like it must've been in our Lord when He looked over Jerusalem, His own people (See?), said, «Jerusalem, Jerusalem, how oft would I have hovered you as a hen would her brood, but you would not.»	Kao što mora da je bilo u našem Gospodu kada je gledao na Jeruzalem, Svoj vlastiti narod (Razumijete?), da je rekao: «Jeruzaleme, Jeruzaleme, koliko sam se puta htio nadviti nad vama, kao što bi kvočka nad svojim pilićima, ali ne htjedoste.»

The King James Version (KJV), also referred to as the King James Bible (KJB) or the Authorized Version (AV), is an Early Modern English translation of the Christian Bible, commissioned by King James VI and I and published in 1611. This translation holds immense cultural significance within English literature and language, shaping literary expression for centuries. Notably, the KJV is in the public domain, allowing for widespread distribution and utilization. The KJV has been automatically aligned using chapter and verse numbers. This method of alignment bypasses the pitfalls of imperfect text alignment generated by standard algorithms, which can introduce errors into the translation pipeline, ultimately compromising translation quality (Bibleverse). Furthermore, the decision to use the KJV as a foundational text was informed by its exclusive citation within the writings of William Branham. This deliberate choice underscores the importance of linguistic and contextual consistency in the study and translation of religious texts. Two randomly extracted excerpts from this aligned corpus are provided in Table 3.

Table 3. Selected excerpts from the parallel corpus of the Bible (Bib_{ver})

Who art thou that judgest another man's servant? To his own master he standeth or falleth. Yea, he shall be holden up: for God is able to make him stand.	Tko si ti da sudiš tuđega slugu? Svome gospodaru on stoji ili pada. A stajat će, jer je moćan Bog održati ga.
The Lord GOD hath sworn by his holiness, that, lo, the days shall come upon you, that he will take you away with hooks, and your posterity with fishhooks.	Gospodin BOG se zakleo svojom svetošću da će na vas, evo, doći dani kada će vas odvlačiti kukama, a potomstvo udicama.

Both procedures were straightforward for distinct reasons: the first due to translations being initially generated using appropriate Computer-Assisted Translation (CAT) tools, and the second owing to the sequential numbering of verses within the text. The descriptions of the corpora are provided in Table 4.

In addition to these efforts, a new corpus was compiled by merging the two existing corpora, taking into consideration both levels of alignment. This resulted in

the creation of the MixSent corpus for the sentence-aligned William Branham corpus and the MixPar corpus for the paragraph-aligned William Branham corpus, as described in Table 5.

A notable observation from the compiled corpora is that the number of sentences is slightly higher on the Croatian side. Conversely, all other statistics listed in Table 4 demonstrate an average increase of approximately 16% on the English side of the corpora. Furthermore, expectedly, the Croatian side of the corpora exhibits a noteworthy disparity, featuring between 50-67% more word types and between 20-35% more lemmas compared to the English side, as illustrated in Table 5. While acknowledging the possibility of errors in the automatic lemmatization process, these differences highlight the linguistic nuances and complexities inherent in the language under study.

Table 4. Corpora statistics

	# of pairs	Tokens		Words		# of sentences		Avg segment length	
		En	*Cro*	*En*	*Cro*	*En*	*Cro*	*En*	*Cro*
BibVer	24.996	947.265	776.324	790.943	620.154	35.074	37.234	23	17
BTSents	63.728	1.312.984	1.146.871	1.046.581	925.172	84.049	84.549	12	11
BTPars	36.151								

Table 5. Corpora lexicon sizes

Corpus	# of pairs	Word types		Token-type ratio		Lemmas	
		En	*Cro*	*En*	*Cro*	*En*	*Cro*
BibleVer	24.996	14.160	43.154	67	18	10.318	15.898
BTSents	63.728	24.389	50.076	54	23	15.001	18.742
BTPars	36.151						
MixSent	88724	31.543	74.047	68	25	20.663	26.932
MixPar	61147						

3.2. System and training

In this paper, we employ AutoML Translate,[1] a Google Cloud AI product designed for tailoring NMT engines to specific industries and domains. The AutoML Translation framework uses transfer learning and neural architecture search to develop

1 See https://cloud.google.com/translate/automl/docs for official Google documentation.

new models based on existing NMT models. Notably, it builds upon the Google NMT (GNMT) system, which is a sequence-to-sequence neural machine translation system featuring a deep LSTM network (Chen et al. 2018; Wu et al. 2016) as the baseline model.

This framework is particularly adept at constructing domain-specific customized models using input datasets from the target domain. It excels in not only adapting to specific industry requirements but also in its ability to generalize effectively across various tasks. By leveraging transfer learning and advanced neural network architectures, such systems offer a robust solution for developing and deploying tailored NMT models that address the unique linguistic challenges and nuances present in specific domains.

The dataset is divided into training, development, and test sets. If the dataset contains fewer than 100,000 sentence pairs, AutoML Translation automatically allocates 80% of the dataset for training, 10% for validation, and 10% for testing. The training set is the data the model "sees" during training and is used to learn the parameters of the model, namely the weights of the connections between nodes of the neural network. The validation set, also known as the "dev" set, is utilized to select the best model generated (by evaluating performance on the validation set) and to adjust the model's hyperparameters accordingly. Employing a separate dataset for fine-tuning, the model structure enhances the model's ability to generalize effectively beyond the training data. The performance exhibited by the model on the test set provides valuable insight into its expected performance on real-world data.

3.3. Evaluation

For the evaluation of models, we utilize a well-established metric called BLEU (Doddington 2002). BLEU is a precision-based metric that quantifies lexical similarity on a 0-1 scale, where 0 represents the lowest score. BLEU compares translation n-grams with n-grams from a reference translation and counts the number of matches at the sentence level, but this count is clipped to the maximum n-gram count found in the reference. These sentence counts are then aggregated over the entire test set. The matches are independent of word position within the sentence. Adequacy is reflected in word precision, while fluency is reflected in n-gram precision. Translations that are significantly shorter than the reference are penalized using a brevity penalty with an exponential decay. It is advisable not to compare BLEU scores across different corpora and languages. However, Table 6 provides an interpretation of BLEU scores that can serve as a rough guideline.

Table 6. BLEU score interpretation[2]

BLEU Score	Interpretation
< 10	Almost useless
10 - 19	Hard to get the gist
20 - 29	The gist is clear, but has significant grammatical errors
30 - 40	Understandable to good translations
40 - 50	High quality translations
50 - 60	Very high quality, adequate, and fluent translations
> 60	Quality often better than human

4. Results and Discussion

The evaluation results are presented in Table 7. The base NMT model achieved BLEU scores ranging from twenty to thirty-three. According to the interpretation provided in Table 6, the base model produced translations ranging from understandable to good for William Branham texts but exhibited significant grammatical errors when translating the Bible and the combined test set.

After adapting the base model, the resulting BLEU scores again fell into two categories. Specifically, translations of the Bible were generally categorized as understandable to good, whereas the translations of the other four cases could be considered high-quality.

The results indicate that the NMT base model performed poorest when translating the Bible. Consequently, adapting the model using the Bible training data resulted in the most significant BLEU score improvement. Furthermore, the NMT base model demonstrated better performance in translating pure William Branham texts compared to the combination of Bible and William Branham texts. The least improvement in BLEU score was observed when translating William Branham texts only (+10.47). The remaining three BLEU score improvements were approximately equal, suggesting that combining training data resulted in similar performance gains. However, the highest translation quality was achieved when training solely on William Branham texts with sentence-level alignment (45.1 BLEU).

Additionally, in the merged corpora scenarios, the level of alignment seemed to have less impact compared to homogeneous scenarios, where more finely-grained alignments had a more pronounced effect on quality improvement.

2 Google Cloud. 2024. *Evaluating models*. Available at: https://cloud.google.com/translate/automl/docs/evaluate

Table 7. Performance of trained models

Model	Google NMT	Google AutoML	BLEU score gain/loss
ModelVerse	20.34	36.02	**+15.69**
ModelSent	32.84	**45.1**	+12.26
ModelPar	32.86	43.33	+10.47
ModelMixSent	27.74	40.1	+12.36
ModelMixPar	27.73	39.91	+12.18

Although no formal human evaluation procedure has been conducted, informal conversations with several translators revealed that even the lowest-scoring model was deemed useful. However, there was a consensus that the highest-scoring model was indeed the best among them.

5. Conclusion

The importance of domain customization is best seen with the occurrence of different crisis such as COVID-19, which are especially challenging for MT engines.

In this paper, Google's AutoML Translation service is used to adapt the base English-Croatian Google NMT model to the field of religion. The corpora used in the training procedure is the King James Version of the Bible and its translation by Ivan Vrtarić dating back to 2016 and the selected texts of William Branham and their respective translations extracted from the translation memories provided by a group of translators. The Bible parallel corpus is aligned at the level of verses, while the texts of William Branham are aligned at the level of sentences. Using the corpora described above eliminated the need for applying text alignment algorithms and errors that might occur.

Since the provided corpora differ in genre and alignment level, several models are built by running the training procedure on each individual corpus, but also on their combination. William Branham texts were additionally aligned at a higher level, i.e., level of paragraphs. This resulted in three models trained on individual parallel corpora, and two models trained on their combination. The Bible parallel corpus was aligned at a fixed verse level in all instances. The results show that the best model, measured by the BLEU score, is obtained when training on William Branham texts alone aligned at the sentence level. The results also show that the base NMT model achieved the best score on William Branham's texts.

Even though formal human evaluation is not presented in this paper, translators who continued working on translations of religious texts assessed the custom-tailored systems positively and felt that all these systems produce translations which can be used as a starting point for human post-editing. Moreover, the best scoring system proved indeed the best in their translation practice.

References

Al-Sabbagh, Rania. 2024. "ArzEn-MultiGenre: An Aligned Parallel Dataset of Egyptian Arabic Song Lyrics, Novels, and Subtitles, with English Translations." *Data in Brief*. doi: https://doi.org/10.17632/6k97jty9xg.4

Carlson, Keith; Riddell, Allen; Rockmore, Daniel. 2018. "Evaluating Prose Style Transfer with the Bible." *Royal Society Open Science* 5(10). doi: https://doi.org/10.1098/rsos.171920

Chen, Mia Xu; Firat, Orhan; Bapna, Ankur; Johnson, Melvin; Macherey, Wolfgang; Foster, George; Jones, Llion; Parmar, Niki; Schuster, M.; Chen, Zhifeng; Wu, Yonghui; Hughes, Macduff. 2018. "The Best of Both Worlds: Combining Recent Advances in Neural Machine Translation.". doi: http://arxiv.org/abs/1804.09849

Doddington, George. 2002. "Automatic Evaluation of Machine Translation Quality Using N-Gram Co-Occurrence Statistics." In *Proceedings of the Second International Conference on Human Language Technology Research, HLT '02*. 138–45. doi: https://doi.org/10.3115/1289189.1289273

Koehn, Philipp. 2020. *Neural Machine Translation*. Cambridge University Press.

Koehn, Philipp; Knowles, Rebecca. 2017. "Six Challenges for Neural Machine Translation." doi: http://www.statmt.org/wmt17

Ruopp, Achim. 2020. "Using Contemporary US Government Data to Train Custom MT for COVID-19." In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas*. doi: https://www.cdc.gov

Viswanathan, Aditi; Wang, Varden; Kononova, Antonina. 2019. "Controlling Formality and Style of Machine Translation Output Using AutoML." *Annual International Symposium on Information Management and Big Data*. https://cloud.google.com/translate/automl/docs

Wu, Yonghui; Schuster, Mike; Chen, Zhifeng; Quoc, V. Le, Norouzi, Mohammad; Macherey, Wolfgang; Krikun, Maxim; Cao, Yuan; Gao, Quin; Macherey, Klaus; Klinger, Jeff; Shah, Apurva; Johnson, Melvin; Liu, Xiaobing; Kaiser, Łukas; Gouws, Stephan; Yoshikiyo, Kato; Kudo, Taku; Kazawa, Hideto; Stevens, Keith; Kurian, George; Patil, Nishant; Wang, Wei; Young, Cliff; Smith, Jason; Riesa, Jason; Rudnick, Alex; Vinyals, Oriol; Corrado, Greg; Hughes, Macduff; Deanet, Jeffrey. 2016. "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation." http://arxiv.org/abs/1609.08144.